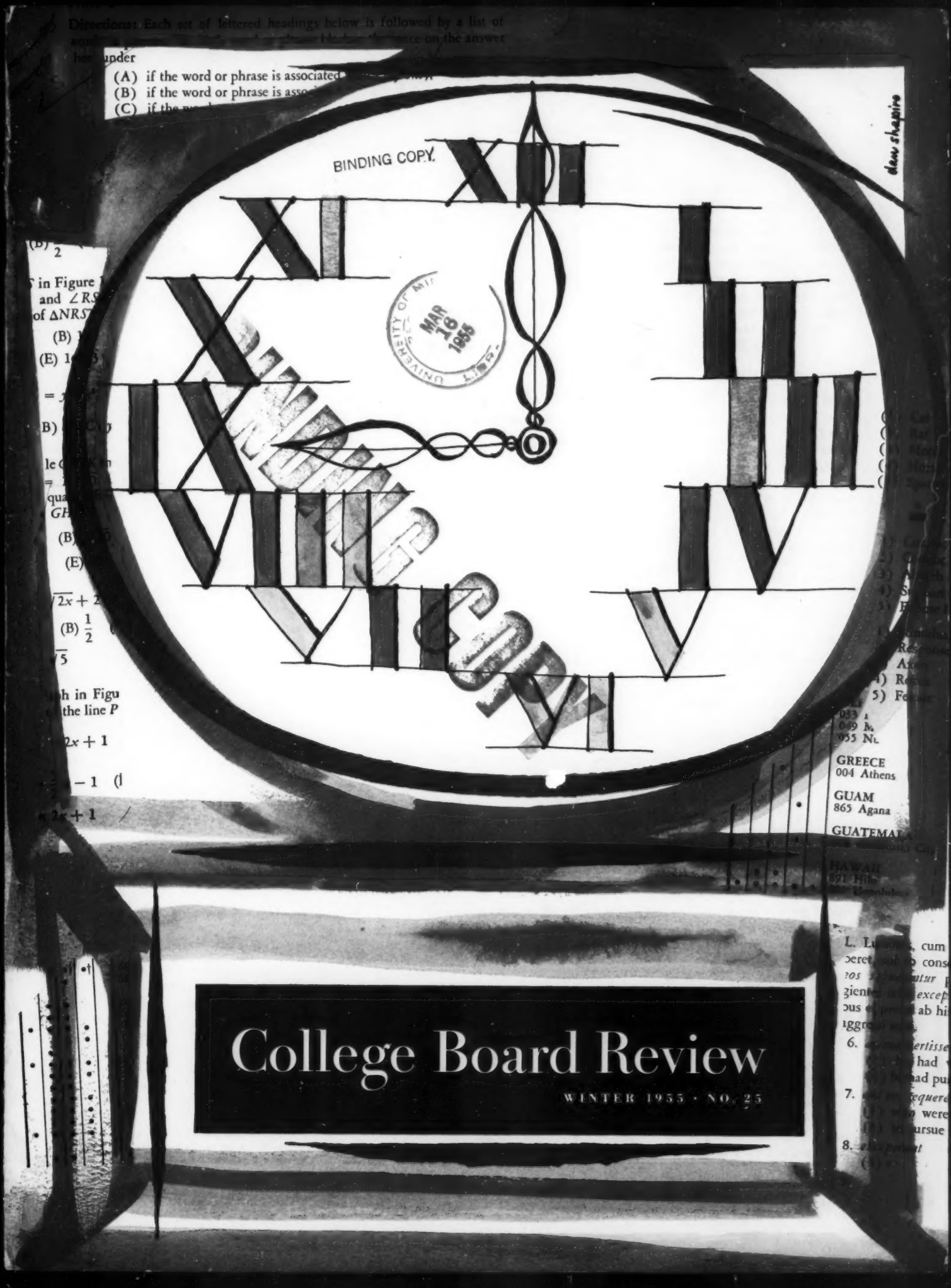

BINDING COPY
MAR 16 1955
UNIVERSITY OF MICH.
dave shapiro
College Board Review
WINTER 1955 · NO. 25

College Board Review subscription: $.50 per year; single copy: $.25. Subscription offices: College Entrance Examination Board c/o Educational Testing Service, P. O. Box 592, Princeton, N. J., or P. O. Box 27896, Los Angeles 27, Calif.

College Board Review

News and Research of the College Entrance Examination Board
published three times a year by the College Entrance Examination Board
425 West 117th Street, New York 27, N. Y.

Contents

Illustrations: The cover by Dan Shapiro gives time the prominence it deserves in all phases of College Board test construction, administration, and score reporting, and the hour the importance it assumes to candidates throughout the world as they start the Scholastic Aptitude Test at 9:00 a.m. *sharp.* Other aspects of college admission are presented in pen and ink drawings by Stanley Wyatt on pages 3, 6, 8, 9, 13, 14, 15, 16, 17, and 22.

NEWS OF THE COLLEGE BOARD

Spring meeting

April 6: A morning symposium and an afternoon business meeting devoted to the discussion of present and possible future activities of the College Board have been scheduled for the April 6 spring meeting.

Among the important items to be considered in the afternoon will be the question of whether or not the General Composition Test (see page 2) should be continued in the testing program. Other business to be considered include a proposal to expand the Executive Committee so that the areas of interest within the Board membership may be more adequately represented by the committee, presentation of a plan for the first administration of the Advanced Placement Tests in the spring of 1956, and the reports and recommendations of committees.

The morning symposium will deal with problems arising from emerging patterns of admission which are tending to complicate the transition from school to college.

Forthcoming publications

Director's Report: A full account of College Board activities in the 1953-54 academic year, including extensive data on the number and distribution of candidates and test score reports, appear in the 53rd *Report of the Director*, which will be sent to subscribers in March.

In the *Report* Director Frank H. Bowles describes and appraises such significant steps as the Board's establishment of a scholarship service (see page 22), a National Science Foundation study of methods to reduce the loss of capable science students in the transition from school to college, and the adoption during the year of more rigorous Board membership requirements.

College Admissions, No. 2: A special publication which will cover the complete proceedings of the second College Board Colloquium on College Admissions is scheduled for March. *College Admissions, No. 2* reproduces the four-day series of classes at which the admissions officers of 90 Board member colleges last October discussed the effects of social, intellectual, personal, and financial factors on college attendance. The cost of the book is $3.50.

Standing committees

Vacancies filled: Appointments recently announced by Provost Samuel T. Arnold of Brown University, Chairman of the College Board, have completed the membership of Board standing committees for the current year.

The places of retiring members of the Committee on Research and Development were taken by Professor Irving Lorge, Teachers College, Columbia University, and Professor Quinn McNemar, Stanford University. Retiring committee members are Professor Andrew H. MacPhail, Brown University, and Professor James A. McClintock, Drew University.

New members of the Committee on Nominations are Dr. Finla G. Crawford, vice chancellor, Syracuse University; Dr. William H. Cornog, president, Central High School, Philadelphia, Pennsylvania; Mrs. Ruth W. Crawford, director of admission, Smith College; and Hubert S. Shaw, director of admissions, Bowdoin College. They succeed Dr. Allan V. Heely, headmaster, Lawrenceville School, Lawrenceville, New Jersey; Dr. Wilson Parkhill, headmaster, Collegiate School, New York, New York; Professor B. Alden Thresher, director of admissions, Massachusetts Institute of Technology; and Annie C. Whiteside, registrar, Randolph-Macon Woman's College.

The largest number of new member appointments was in the Committee on Examinations, which now includes Professor Frederick B. Agard, Cornell University; Edwin C. Douglas, assistant to the headmaster, Taft School, Watertown, Connecticut; Harold Howe, II, principal, Walnut Hills High School, Cincinnati, Ohio; Ruth Jenkins, headmistress, Annie Wright Seminary, Tacoma, Washington; Professor Arthur E. Jensen, Dartmouth College; Professor Thomas C. Mendenhall, Yale University; and Professor Denham Sutcliffe, Kenyon College. Retiring members of the committee are Dr. Robert G. Andree, headmaster, Brookline (Massachusetts) High School; Professor Harry D. Berg, Michigan State College; Dr. William H. Cornog, president, Central High School, Philadelphia, Pennsylvania; C. William Edwards, director of admission, Princeton University; Dean Everett L. Hunt, professor of English, Swarthmore College; Allegra Maynard, headmistress, Madeira School, Greenway, Virginia; and Donald D. Walsh, director of senior studies and Spanish department head, Choate School, Wallingford, Connecticut.

Visiting representative

To see urban colleges: T. Leslie MacMitchell, former assistant dean of admissions at New York University, has joined the staff of the College Board on a temporary basis to visit member and non-member urban colleges during the spring months. The purpose of his visits will be to collect information on the admissions procedures and problems of degree-granting afternoon and evening study programs.

Mr. MacMitchell recently resigned from the Brooklyn Dodgers organization in which he served for one year as executive assistant to the president.

Should the General Composition Test be continued?

The General Composition Test, grown familiar during the past four years as an idea, then as an experiment, is scrutinized for the first time in the two articles which follow in terms of its performance as a regular College Board test.

There are two articles because at least two positions may be taken on the test and the results of its first administration. The positions are far apart, based on differences in theory which could have been advanced just as firmly five years ago before there was a GCT, and before the quest for "a test of writing ability" could be expressed as a debate on specific testing procedures or statistical interpretations.

In simplest and most extreme language, it may be said that those who espouse one or the other of these two viewpoints do not see the same thing when they look at the GCT. They not only see the other side of the same coin; the other side is stamped with a different value in another language.

Presented with any test which is to be introduced as a college entrance examination, the spokesman for one point of view will claim that the test should measure the kinds of things that are taught in school, and should produce the best possible prediction of the candidate's success in college. This can be done best, he will say, by tests that are "reliable," that is, may be scored so consistently that the same score always represents very close to the same degree of achievement, and that they are "valid," that is, forecast how well or poorly the student will do academically in college.

The test is, first and foremost, a test, and the ways in which its reliability and validity characteristics affect its use for college admissions purposes are considered of essential importance. Whether it is an objective test, or an essay test like the GCT, it must be effective as a test which measures consistently and accurately.

The opposite opinion denies that any weakness the essay test may have as an instrument of selection should override its superior value as an aid to instruction. Proponents of this persuasion point to the interest and enthusiasm of schools which have used the GCT and the unexpectedly large number of candidates who took the test the first time it was scheduled by the College Board. If teachers armed with a good essay test are thereby enabled to improve the general state of English composition and provide the colleges with a more literate candidate, it is argued, the application of validity and reliability standards to select students on the grounds of precise differences in writing ability becomes progressively less important.

Moreover, it is unfair and unfeeling, in a most vital sense, to measure an essay test according to objective test standards and only by them. The best test of writing, it is proposed, is writing itself; the best judges of English composition are those who teach it and who are selected by the College Board to work together on this test. They do not expect to see a perfect essay test any sooner than a purple cow, but they do believe the GCT is a good test which will encourage improvements in English teaching.

The question which the College Board will face at its meeting this April is whether the unique educational virtues claimed for the GCT compensate for an apparent loss in reliability which is conspicuous in contrast to other College Board tests. If the Board accepts the GCT as a permanent feature of its admissions test battery, it must also accept the risks inherent in a test which does not behave as consistently as the other tests in the program.

The recommendation of the officers and staff of the Board and the Educational Testing Service is that the GCT be offered at least one more year after this May, when the two-year trial period will expire, and that certain studies be conducted during that time to complete general understanding of the test. These would probably include an attempt to appraise the educational value of the test, a comparison of the performance of the GCT with that of the Board's objective English Composition Test, the exploration of possible ways to improve GCT reliability, and an attempt to adjust scores for recognizable differences in the leniency or strictness with which certain readers habitually grade essays.

The test fails as an entrance examination

BY RICHARD PEARSON

The General Composition Test had its genesis in a report of a special College Board Committee on English Testing which met in 1949-50 and expressed its concern that "when the essay type test was abandoned [in the Board's program] there ceased to be any testing of certain crucially important educational values." The committee "dreaded that unless these values could be re-emphasized, or possibly even rediscovered, the effect on American

education would be most unfortunate."

The result of this concern was a two-hour test of writing ability which requires the student to read background materials on the topic presented, outline an essay on the topic, write the essay, and then summarize its theme briefly. This is the test which was offered last May for the first time in the regular College Board program and which is scheduled for a second, and possibly final, administration in May, 1955. The future of the GCT will depend upon action to be taken by the Board at its spring meeting on April 6, when it will become necessary to reconsider the two-year trial period approved by the Board in 1953.

From tryout to regular test

The events which led up to the introduction of the test included experimental administrations of tryout forms to 800 students from 29 public and independent schools in the spring of 1951 and to 800 freshmen at four colleges in the fall of the following year. The experiments developed procedures for the selection and presentation of essay topics and for the reading of students' papers so that their scores would be reported in terms which would be meaningful to admissions officers. During this time, in response to inquiries about the test, copies of the tryout forms were also made available at cost to interested schools for their administration, scoring, and use. In the spring of 1952 more than 300 schools ordered more than 8,000 copies of the test. The following spring, when a second form of the test was offered, about 700 schools ordered more than 14,000 copies.

Among the factors to be considered by the Board in April and by the appropriate Board committees which early in March will consider whether or not to recommend continuation of the test is a sizable fund of information gleaned from the first regular administration and studied by the Educational Testing Service. The comments which follow summarize those studies and represent the ETS interpretation of data made available by that one administration.

A total of 4,272 candidates took the GCT in May 1954 of whom 96% were preliminary or junior year candidates,

86% were students from independent schools, and 64% were girls. A special study of the candidates tested by the ETS office in Princeton indicates that 30 colleges which received GCT score reports for 20 or more candidates accounted for a total of 2,610 reports. Of these, 1,630 score reports, or 62%, were sent to the women's colleges comprising the Seven College Conference group,[1] although apparently not as a requirement for admission. Except for the Seven Colleges and possibly Connecticut College, which received 121 score reports, no college appears to have requested scores on this test generally for junior year applicants. A total of 80 secondary schools with 20 or more candidates, of which 74 schools were independent and six were public, requested and received their GCT scores. These schools were represented by 3,180 reports, or 74% of the students taking the test.

At the same time that the GCT was offered in the regular May series, alternate forms of the test were made available to the colleges at nominal cost through the Board's regular placement testing program. During 1953-54 two colleges ordered the GCT for placement use and a total of 143 tests were administered. In the current academic year one college has ordered 44 copies. It should be noted that the offering of this test in the placement program did not include reading of the essays by the College Board.

Reading of the May 1954 examination employed careful techniques developed during the two experimental reading conferences to secure a common understanding of grading standards. An intensive period of discussion, training, and practice reading was held at the beginning and the training process was continued throughout the conference—training not to impose fixed standards of excellence, but to secure full cooperation from each reader in the definition of these standards. Each essay was read twice by two different readers and day-to-day records of the progress of the reading were maintained and evaluated. The reading conference, held at Haverford College in the last week of June, was described as follows by Earle G. Eley, the chief reader and guiding spirit of the test since its beginning:

"One question was whether 4,300 essays could be read in the allotted time. This question was answered in the affirmative. The average time for two readings of an essay plus a third reading when necessary during the two earlier experiments had been 27½ minutes. Thus it was estimated about 30 minutes per essay for the Haverford Conference. However, neither of the two previous experiments had allowed for the actual reading after completion of training to run for more than three days. There was some indication that the speed of reading might increase significantly if a conference were extended to allow for as many as five days of uninterrupted reading. This actually proved to be the case. It was estimated that during the time available at Haverford, as many as 6,000 essays could have been read.

"Another question—could we develop a spirit of unity among the readers—it was also possible to answer in the affirmative. Although we were not able to hold the pre-dinner meetings

[1] Barnard, Bryn Mawr, Mount Holyoke, Radcliffe, Smith, Vassar, and Wellesley Colleges

that had during the two experimental conferences been so large a factor in developing rapport, we were pleased to find that even so large a group as 52 readers by the second or third day had taken on a quality of enthusiasm which made for a spirit of unity in the conference. The readers pronounced a verdict in favor of the test which was almost unanimous. Despite the grueling task of reading essays for six hours a day, day after day, readers continued to be interested in the problems of judgment and continued to discuss among themselves problems of reading during the morning and afternoon coffee breaks.

"In the earlier conferences dealing as we did with no more than 800 cases, two clerical workers and three administrative workers using a system of hand entries both for computing final grades and for computing our continuous statistical check on reader agreement proved feasible enough. At the Haverford Conference, it became clear that when the number of candidates goes into the thousands a system of hand entries is not feasible even though we did manage at this conference to keep up after a fashion."

The possibility of undertaking this clerical work by means of IBM tabulating equipment has subsequently been studied by ETS and it seems likely that in the next reading conference the use of this equipment can be counted on to provide the following clerical services: (1) the automatic resolution of grades assigned by the first and second readers where differences between the two readings fall within pre-determined limits; (2) the identification of papers where disagreements between the first two readers are sufficiently great to require a third reader; (3) summary information, reader by reader, on the work of a given day which will be available by the morning of the following day for use in training. It appears that as many as 6,000 essays could be handled with a combination of the same number of readers and an improvement in our clerical procedures. It is doubtful, because of the difficulty of obtaining readers before the close of school, that the scores can be reported in substantially less time than the two months required for last May's test. It also seems that an increase to 10,000 or 15,000 candidates could be scored by a proportionate increase in the number of readers and adaptation of the tabulating equipment. Whether a greater number could be handled with these procedures is uncertain.

The test itself

The General Composition Test is a two-hour test of writing ability, which will be offered only at the May series. Candidates taking this test may take only one other Achievement Test at this series.

The essay topic for this test is a general problem of wide interest. The student is given reading materials for background on the topic presented. After reading the background materials, the student is asked to prepare an outline for his essay, to write the essay, and to summarize its theme briefly.

Two problems which have been used as topics are: "Should women be given the same educational and professional opportunities as men?" and "Is there a conflict between science and human values?"

The preliminary instructions to the student follow:

Directions: Today you will be given a problem stated in general terms and a set of reading materials dealing with one case in which this problem has come up. After you have read the materials carefully, you will be asked to write an essay stating as well as you can what you think about the problem.

Your essay will be judged impartially by qualified readers according to five different qualities. These five qualities are:

1. *Mechanics.* Your ability to use correct forms of grammar, to punctuate and spell correctly.
2. *Style.* Your ability to state your ideas clearly and effectively in language which is appropriate to your essay.
3. *Organization.* Your ability to arrange the parts of your essay in such a way as to make your ideas and purpose clear to your readers.
4. *Reasoning.* Your ability to support your conclusions with evidence and to reason in such a way as to convince your readers of the soundness of your conclusions.
5. *Content.* Your ability to supplement the reading materials given with material drawn from your general reading, your school studies, and your own life.

You have two hours for the complete examination. Be sure to read all the material and all parts of the writing assignment before you begin to write. It is suggested that you spend about twenty minutes reading the materials. In the hour and forty minutes remaining you should plan and write your essay.

College Board Tests, 1954-55, The College Entrance Examination Board Bulletin of Information, p. 35.

The question of reliability

An important measurement characteristic of the GCT is the extent to which the readers agree on the assignment of scores to a given paper, that is whether they agree on what constitutes a high-scoring or a low-scoring paper. The question is not whether the readers agree, for it may be assumed that there will be some disagreement, but rather how much they disagree and if this disagreement falls within tolerable limits.

The GCT score scale reported to schools and colleges ranges from 1 (low) to 4 (high) points. A score of "1" assigned by a reader indicates the reader's belief that the student is unprepared to do college composition. Similarly, a "2" indicates the reader's judgment that the student needs additional training, although he may be able to do college work, a "3" that he

could do satisfactory work but would benefit materially from an introductory writing course, and a "4" that he wrote well enough so that he might be excused from introductory writing courses. These were the "labels" used in the reading process; they do not imply any inherent recommendations for selection of placement. The scores are reported in the categories of Mechanics, Style, Organization, Reasoning, and Content, each paper receiving a score in each category. During the early experimentation with the GCT, these five category scores were added for each student and a composite GCT score was derived. Since the Board's Committee on Examinations decided not to report composite scores for the May 1954 administration, we have not obtained statistics for the composite in all of the studies which follow.

In an attempt to determine the extent to which the readers agreed in scoring, Frances Swineford of ETS traced a random sample of 370 papers through the reading process.[2] Selected to be typical of all papers read last summer and not differentiated from the others at any point during the reading, these essays received the regular two independent readings. This step is carefully controlled by an elaborate routing system to insure that each reader reads "against" every other reader approximately to the same extent throughout the conference. It should be observed that the readers use a more refined score scale at this point than is used to report final scores. Permission is given to use "borderline" categories when, in the reader's judgment, a paper deserves a "low 4" or a "high 2." The effect of this is to expand the original 4-point scale to a 10-point scale which is subsequently contracted back to a 4-point scale when final scores are assigned.

It will be helpful now to refer to the 370 papers in Dr. Swineford's study after the scores of the first two independent readings were obtained, using the 10-point scale. To simplify the presentation, we have selected only two of the five categories, Style and Organization. The Style category was the one in which greatest agreement among the readers was secured; the Organization category was among the categories most difficult to read. Dr. Swineford's comparison of the first and second readings of these categories for the 370 papers in the sample yielded the following information:

Number of papers	*Style*	*Organization*
With complete agreement between two readers	92	78
With disagreement of:		
One point	100	90
Two points	91	81
Three points	59	65
Four points	22	35
Five points	4	10
Six points	2	10
Seven points	—	1
Eight points	—	—
Nine points	—	—
Total number of papers	370	370

The product-moment correlations summarizing the amount of agreement between first and second readings are .57 for Style and .43 for Organization. Corresponding values for the other categories are .53 for Mechanics, .42 for Reasoning, and .44 for Content.

Reconciling disagreements

The second step in the reading process was a reconciliation of some of these disagreements between first and second readers. This reconciliation was made solely with reference to differences in first and second reading scores; no attempt was made to reread any papers in this step. The procedure is as follows:

All one-point disagreements are resolved by (1) accepting the integral score as the final score, in cases where one reader assigns an integral score and the other reader assigns an adjacent borderline score; and (2) determining the final integral score by an inspection of final scores assigned to other categories when both readers assign scores on the same borderline but indicating adjacent integral scores. An example of the first of these is a first reading score of "3" and a second reading score of "low 3." The final score would be the integral score "3." An example of the second of these is a first reading Style score of "low 3" and a second reading Style score of "high 2." The integral score of "2" or "3" would be assigned the paper as a final Style score, depending upon the preponderance of "2's" and "3's" among Mechanics, Organization, Reasoning, and Content for that paper.

All two-point disagreements are resolved by (1) accepting the integral score as the final score, in cases where one reader assigns an integral score and the other reader assigns a borderline score which is adjacent to the next higher or lower integral score; and (2) accepting as the final score that integral which falls between two borderline scores, in cases where one reader uses the upper border and the other reader uses the lower border of the same integral score. An example of the first of these is a "2" and a "low 3"; a final score of "2" would be assigned this paper, without further reading, despite the fact that the second reader indicated a preference for a "3." An example of the second of these is a "high 3" and a "low 3"; the final score in this case is "3."

Some three-point disagreements are resolved when two borderline ratings were assigned by the two readers and when one of the borderlines indicated a preference for a given integral score and the other indicated a preference for the next higher or lower integral score. That is, a "high 2" and a "high 1" would be assigned a final score of "2," despite the fact that the second reader indicated a preference for a "1."

The effect on the 370-case sample of this reconciliation process is summarized below:

Number of papers	*Style*	*Organization*
With complete agreement between two readers	92	78
With disagreements which have been resolved by adjusting scores of two readers	208	179
With remaining disagreements of:		
Three points	42	57
Four points	22	35
Five points	4	10
Six points	2	10
Seven points	—	1
Eight points	—	—
Nine points	—	—
Total number of papers	370	370

The final step in the reading process

[2] "Reading Reliability of the General Composition Test, Form H," *Statistical Report* No. 54-34, October 1954, Educational Testing Service, Princeton, N. J.

is to refer the papers with disagreements remaining between first and second readers to a third reader. Normally a third reading is sufficient to resolve these discrepancies and it was for the remaining 70 Style and 113 Organization disagreements in the sample.

In other words, in their reading of 370 essays and scoring them on a 10-point scale, the first and second readers were in complete agreement on the Style score in 25% of the cases and on the Organization score in 21% of the papers. If the definition of "agreement" is broadened to accommodate the reconciliation of scores showing one or two, and in a few cases, three points of difference, the agreement for Style rises to 81% and for Organization to 69%. The agreement reached 100% in both cases with resolution by a third reading.

Strict or lenient?

Which two readers graded a paper was also found to have an important bearing on the reported scores. This factor was recognized in planning for the conference because it seemed likely that a paper scored by the strictest and the most lenient of readers might have to be referred to a third reader, while scorings by two readers who were both strict, both lenient, or both moderate would by reason of that pairing, rather than the quality of the student's writing, be in close or complete agreement. Great care was taken in the selection of experienced readers and in the training part of the conference to avoid variations in scoring caused by the readers and Dr. Swineford sought to discover how successful these precautions had been. Twenty-four papers were selected at random, none from the sample of 370 just described, and each was read by 15 readers according to a most detailed routing design which protected the integrity of the study. Since each paper was read by 15 readers, there were 105 instances for each paper where different pairs of readers were involved. Thus, it is possible to know how the papers would have been scored by one pair of readers, then by another pair, and so on through 105 pairs.

Taking the first of two extreme papers for illustrative purposes—Paper A, which was easy to score—we see below what actions would have been taken in the 105 cases, at the step in the reading process where scores of first and second readers are compared. The scores are given in the five categories, M (Mechanics), S (Style), O (Organization), R (Reasoning), and C (Content).

A paper scored by a strict and a lenient reader might need a third reading

Final grade	M	S	O	R	C
4	66	101	65	72	78
3	24	4	10	17	5
2	—	—	—	—	—
1	—	—	—	—	—
Third reader	15	—	30	16	22
Reader pairs	105	105	105	105	105

Paper A would have received a grade of "4" in Mechanics 66 times out of 105, depending solely on which two readers were given the paper during the first and second readings. Twenty-four times out of 105, a final Mechanics grade of "3" would have been assigned to the papers as a result of the first two readings. The total of these, or 90 times out of 105, gives the number of times when agreements between first and second readings were within previously defined limits and resulted in no third reading.

The following actions would have been taken on Paper B, which was difficult to score:

Final grade	M	S	O	R	C
4	2	—	—	—	—
3	28	15	24	36	43
2	37	30	49	12	23
1	1	6	2	6	2
Third reader	37	54	30	51	37
Reader pairs	105	105	105	105	105

The variations in final scores which would result if Paper B were given first and second readings by different pairs of readers are considerable. A final Mechanics score over the entire scale is possible, depending solely upon which two readers were given the paper. Final scores in Style, Organization, Reasoning, and Content over three of the four scale points are possible. And even though Paper B is difficult to score, there is no firm probability that a third reading will be given this paper under existing reading procedures. It must be concluded from this evidence, which is supported by more general statistics from Dr. Swineford's study, that *which* two readers participate in the first and second readings has a decided influence on the reported scores.

The crucial question in this consideration of reader agreement is what would happen if all papers from May 1954 were graded a second time by means of a second reading conference. To what extent would the same papers receive the same scores? A second conference was not held but Dr. Swineford did estimate what the results would have been from data collected in the Haverford reading. Her general conclusion was that "the reading reliabilities for the reported category scores probably range from the high .50's to about .70, when the scores are based on ratings by two or three readers." Mechanics and Style generally showed a greater reader agreement than did Organization, Reasoning and Content. Although a composite GCT score was not obtained at Haverford, it is not likely that the reading reliability of this score would exceed .75.

The difference in practice

The implications of this finding may be seen by posing an operational question facing a school or college official: "If I am considering two students each of whom took the GCT last May, how much of a difference in score between the two should there be before I can say with confidence that the difference *did not* arise because of reader disagreement?" Remembering that GCT scores are reported on a four-point scale, the evidence from Dr. Swineford's study suggests that the statement could not be made with confidence unless the score difference were two points—one less than the maximum possible for any two GCT scores. It appears evident that reader disagreement was a major source of error in the GCT scores reported to schools and colleges and that those of individual students should be used with that in

mind. It is also likely that it would take 10 to 15 readings of each paper to eliminate that source of error, a situation which suggests that while some improvement in reading reliability may be achieved by future refinements in the reading process, there is little hope that these can altogether solve a serious problem.

The question of validity

A second important measurement characteristic of the GCT is the extent to which the scores reported for individual students agree with independent estimates of the students' writing ability. In this case, the question is whether the scores "make sense"—whether they rank the students in a way which would be meaningful to teachers who are presumed to know which of their students are better than others in their writing ability.

The application of this type of an analysis to any test is both comprehensive and ambiguous. It is comprehensive because it considers both the errors arising from reader disagreement and those which result if the students are not challenged equally to turn in their best performance. It is ambiguous because one fallible measure of writing ability (the GCT) is set against another (teacher estimates). Given less than unanimity, the question of where the difference lies inevitably occurs. Such an analysis of the GCT scores was made, however, in an attempt to discover if the GCT is really a valid *measure* of writing ability.[3]

To obtain an independent appraisal of writing ability, Dr. Marjorie Olsen of ETS asked teachers at 11 secondary schools to rate their students on the same five categories used in scoring the GCT, basing their ratings on observations extending over the school year, not on any single factor, and without prior knowledge of the GCT scores. The ratings at nine schools were supplied by the English teacher, and at the other two schools by the English teacher, with one or more teachers of other subjects collaborating. At the same time, junior year English and average grades for these students were obtained from school records as supplementary criteria. A total of 617 students were involved in the study.

The evidence from Dr. Olsen's study on how well the GCT scores predicted the teachers' estimates is contained in the correlations between Mechanics score and Mechanics rating, Style score and Style rating, etc. The averages of these correlations over the 11 schools are as follows:

Ratings:	M	S	O	R	C
GCT (Corresponding scores)	.43	.41	.29	.36	.33

By way of contrast, it may be shown how well the teacher estimates might have been predicted if, instead of selecting the "appropriate" GCT score (i.e., the same category the teacher was rating), the GCT score showing the highest actual relationship were used. The average correlations are as follows:

Ratings:	M	S	O	R	C
GCT (Most predictive scores)	.43	.41	.34*	.37*	.37*

The entries with asterisks indicate cases where "inappropriate" GCT scores showed higher relationships with the criterion than did "appropriate" GCT scores. The Organization rating was better predicted by three other GCT scores (Mechanics, Style, and Reasoning) than it was by the Organization score. The Reasoning rating was slightly better predicted by the Style score than it was by the Reasoning score. The Content rating was better predicted by the Style and Reasoning scores than it was by the Content score. The Style score, in fact, was the best predictor of the five in the case of every rating but Mechanics, where it was second best. The Style and Mechanics scores, it will be remembered, also showed higher reading reliabilities than the other three scores.

Admittedly, some of these differences are small and may possibly have arisen by chance. Nevertheless, Mechanics and Style are generally considered to be quite tangible whereas Organization, Reasoning, and Content are more elusive. Those who have studied these results at ETS consider it significant that Mechanics and Style held up well on this line of inquiry, whereas Organization, Reading, and Content showed evidence of some confusion.

The Scholastic Aptitude Test Verbal scores were also obtained for the students in this study,[4] with the thought that they might be helpful if ambiguities appeared between the GCT scores and the teachers' ratings. When this proved to be the case, the SAT scores were related to the teachers' ratings, and the following average correlations were obtained for the "appropriate" GCT scores:

Ratings:	M	S	O	R	C
SAT-V	.48	.57	.49	.60	.58
GCT (Corresponding scores)	.43	.41	.29	.36	.33

This showed that the SAT, without exception, predicted the ratings better than did the appropriate GCT score. Since the ratings can be predicted, it would seem that their use as criteria in the present analysis is justified.

Correlational evidence not presented in detail here indicates that each GCT score except Mechanics is more closely related to SAT-V than it is to the corresponding teachers' ratings. In the case of Mechanics, however, the GCT score shows a closer relationship with the teachers' ratings.

We have previously considered only the category scores of the GCT which were reported for last May's test, but these can be summed in whole or in part in order to yield composite scores. In her analysis, Dr. Olsen includes a

Richard Pearson is the Educational Testing Service project director of the College Board test program, occupying a post in which he represents each organization to the other and the best interests of both. Mr. Pearson expresses the well-considered opinion of the ETS in reporting its analysis of the General Composition Test.

[3] "The Validity of the College Board General Composition Test," *Statistical Report* No. 55-4, January 1955, Educational Testing Service, Princeton, N. J.

[4] We would have liked to include the Board's Achievement Test in English Composition in this analysis but there was not a sufficient number of candidates also taking this test last May.

sum of Mechanics and Style and a sum of all five category scores, reporting their correlations with a composite of the teachers' ratings as follows:

	Composite Ratings
M + S	.48
M + S + O + R + C	.51

It is apparent from these figures that the Mechanics and Style categories contribute most of the GCT prediction of teachers' ratings. The correlation for the composite which includes only these categories is .48. When Organization, Reasoning, and Content are added in a total composite GCT score, the correlation increases only to .51.

In addition to the teachers' estimates of writing ability, the study obtained the students' junior year English grades and their average grades over all courses for that year. The averages over the 11 schools for correlations between GCT and SAT scores and these grades were:

	M	S	O	R	C	SAT-V
English grade	.43	.48	.36	.43	.39	.67
Average grade	.37	.42	.31	.35	.33	.58

It may be noted that the SAT is the superior predictor for these two criteria and that both GCT and SAT show higher relationships with English Grade than with Average Grade.

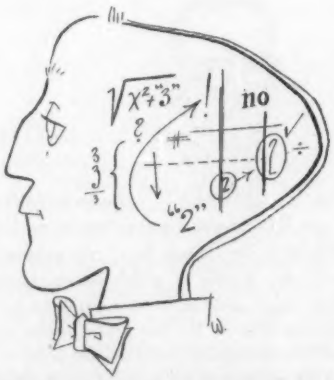
It was also of interest to discover what improvements in prediction would occur if a combination of the GCT and SAT-V were used. Dr. Olsen reports the following correlations:

	Composite Ratings
GCT composite *only*	.51
SAT-V *only*	.65
Simple combination of both	.66
Best-weighted combination of both	.68

This means that if a teacher or an admissions officer assumes each of the tests should have equal weight in the combination, he obtains a correlation of .66, an increase of .01 over the correlation using SAT-V alone. If he uses a "best-weighted" combination, derived uniquely for his school or college using standard statistical techniques, such as was done for each of the 11 schools before averaging them, he may get a correlation of .68, an increase of .03 over the correlation using SAT-V alone. It is evident that the GCT adds very little to the SAT-V.

The most significant finding from both of the foregoing studies is the obvious weakness of the GCT score categories of Organization, Reasoning, and Content. This weakness was demonstrated by the relative lack of correspondence between these scores and the appropriate teachers' ratings. The fact is that readers on the one hand, and teachers on the other do not seem to agree on what constitutes Organization, Reasoning, and Content when they are asked to define these in terms of student performance. This weakness was also apparent when we inquired as to the relative prediction obtained from a composite GCT score derived from Mechanics and Style as compared with a composite GCT score derived from Mechanics, Style, Organization, Reasoning, and Content. When we pursued this line of inquiry we found that prediction from a combination of Mechanics and Style was almost as good as prediction from a combination of all five category scores. Further, this weakness is consistent with the results of the reading reliability work where it was found that the readers showed a greater tendency to disagree on the Organization, Reasoning, and Content categories than they did on the Mechanics and Style categories. The weight of this evidence indicates unmistakably that the GCT is largely a measure of Mechanics and Style and cannot presume to be an effective measure of Organization, Reasoning, and Content.

The question which may be asked in consideration of the data presented above is whether the GCT, alone or in combination with the SAT-V, is a useful predictor for college guidance or admissions. Let us assume that a college teacher wished to place the 617 students from this study into four classes: (1) the best writers; (2) the next-best writers; (3) the next-to-the-poorest writers; and (4) the poorest writers. Assume further that he would like to use, as his criterion of writing ability, the school teachers' ratings from this study, because they are based on day-to-day observations of the students' work. He does not have these ratings but can estimate them from either the GCT total scores, or the SAT-V scores, or a simple combination of the two. How many students would be slightly or seriously misclassified under each alternative?



If he uses a weighted combination . . .

Using the scores

If the college man used the GCT total alone, he would be "right" in 269 of the 617 cases; these students would enter a class indicated by their school teachers' ratings. He would be "somewhat wrong" in 298 cases; these students would be misplaced by one class from that indicated by the ratings. He would be "badly wrong" in 50 cases; these students would be misplaced by two or three classes, from that indicated by the ratings.

If our college teacher used the SAT-V alone, he would be "right" in 297 cases, "somewhat wrong" in 289 cases, and "badly wrong" in 31 cases. For the combination of GCT and SAT-V, he would be "right" in 307 cases, "somewhat wrong" in 272 cases, and "badly wrong" in 38 cases.

Two conclusions can be drawn from this enumeration of students from Dr. Olsen's study. The first is that use of the SAT-V, rather than the GCT, would result in more students properly placed and fewer students improperly placed in classes sectioned according to this criterion of writing ability. The second conclusion[6] is that use of a simple

[6]The reader may wonder why the "best-weighted" combination was not used in this comparison. The reason is that there are technical grounds for suspecting that the procedure introduces a spuriously high degree of correlation, which would not be maintained if the "best-weights" were applied to other groups of students. In the absence of evidence to the contrary, a more conservative position has been taken.

A concern for the individual candidate

combination of SAT-V and GCT, rather than SAT-V alone, would result in a slightly greater number of properly placed students but might also result in increasing the number of badly misplaced students.

The basic issues

The writer is unable to close this report without commenting briefly on some of the issues raised by the inclusion in the Board program of an essay test of writing ability. The first of these has to do with the stringency with which one interprets the available evidence dealing with measurement characteristics of the GCT. The concern of ETS with reader agreement is not academic nor is it an attempt to harass a hard-working group of able readers. Our interpretation of this evidence is related directly to a concern for the individual candidate. From the time he registers for the Board tests until his scores are released to colleges and schools, we feel obliged to impose standards of near-perfect accuracy; the intermediate processes are, or should be, under rather complete control.

If we are asked to attest to the accuracy of reported scores, as we frequently are by candidates, schools, or colleges, we welcome the assurances of procedures which, when repeated, yield essentially the same results. Further, as test people, we know from experience that high consistency in assigning scores to individual candidates is an attainable goal when there is general understanding concerning the task the candidate is asked to perform and what constitutes satisfactory or unsatisfactory performance. Anything less than high consistency of scoring implies that we have not attained this general understanding and hence are operating in the realm of chance, rather than design.

A second issue raised by the GCT is its effect on instruction in the secondary schools. The present report does not present evidence bearing directly on this important issue. It is likely that carefully conceived research could contribute much helpful information concerning ways and means of furthering good teaching practices. It would be a mistake, however, to assume that educational standards and measurement standards are unrelated. To the extent that the results of our measurement are used in education, we run grave risks if we sacrifice standards of one in order to further standards of the other. Any research on a test's influence on instruction should take the test's measurement characteristics into account.

A third issue is the use which will be made of the GCT scores at schools and colleges. If one accepts the interpretation of available evidence, this test is one of moderately high validity but low reliability. This combination of measurement characteristics sharply limits the use of the scores with an individual student. We can be reasonably sure that any group statistics, such as a class average, will be a dependable measure of group performance on the test, since these statistics are less sensitive to unreliability than are the individual scores. We cannot, however, treat the scores of individual students with the same degree of confidence since these bear the brunt of any unreliability in the test. The analogy here is with group statistics of the incidence of a major disease. Medical science can predict with considerable accuracy the incidence of polio or heart disease in large segments of the population. It cannot predict with any certainty whether an individual will contract the disease. For this reason, we believe that any future administration of this test in the Board's regular series should be accompanied by adequate safeguards to minimize any possible misuse of the scores.

And finally, even though we are questioning the effectiveness of the GCT as a measure of writing ability (or perhaps because of this), we recognize the urgency to continue the research efforts of teachers and measurement people alike. We do not know whether progress will be achieved by refinements of this test or whether new lines of inquiry will be more fruitful. We do know that the Board and its committees will continue to give careful consideration to the question of measurement of writing ability.

The test satisfies an educational need

BY EARLE G. ELEY

The present status of the General Composition Test is the result of cooperative endeavor. Since I became associated with this test some four years ago, I have seen it develop through the efforts of 40 or 50 teachers of English and through the serious contributions of a great many more people who, though not teachers of English, are devoted to the idea that writing should be tested by *writing*.

There was a time when no one in his right mind would have dreamed of suggesting that the ability to write be tested by any means other than by having the individual write. But the time came during the twenties and thirties when educational psychologists began to experiment with indirect measures of writing made up of objectively scored items which they hoped would test validly the ability to write. This experimentation arose from the concern about the inconsistency, or more technically, unreliability, with which readers were found to be able to grade essays.

As early as 1912 Starch and Elliott pointed out that common school teachers disagreed in their scoring of the

same essay by as many as 35 to 40 percentage points.[1] These findings were reproduced over and over by other investigators during the twenties. The lack of objectivity in scoring essays constituted a major problem, since most testing at that time was done by essay. It is not difficult to understand that, given the temper of American educational psychology of the period, the efforts to solve this problem should have been directed not to the improvement of essay tests, but to the development of objectively scored tests of writing competence. The substitution of such "new-type" tests for the essay was argued, not on the grounds of validity, but rather on the grounds that essays, however valid, could not be read reliably. The objective tests, it was felt, not only could achieve extremely high statistical reliability, but also were economical to administer, and provided a solution to the problem of growing numbers of candidates. The pioneers of this movement validated their tests by going to the actual human operations which they intended to measure. Thus, in the case of writing, they went to essays.

Product vs. process

It is ironical, but understandable, that the objective test movement should have developed, by now, so intricate a maze of testing procedures that frequently at least a part of the validation of one objectively scored test is made by judging it against other objectively scored tests. This procedure may be illustrated historically by the development of intelligence tests. Mr. Pearson's article, our companion on these pages, affords an example of this modern practice: the criterion for writing which seems to be most highly regarded is the Verbal score of the Scholastic Aptitude Test, an objectively scored measure.

The argument of Mr. Pearson's article is a comparatively simple one. It begins with a description of the GCT and then proceeds to an analysis of the test's reliability and validity. The conclusion drawn is that the GCT is not a satisfactory measure of writing competence. Now the problems of reader reliability and validity posed by the GCT are no different in nature from those posed by any free essay. By structuring the essay question to create a common starting point for both writers and readers, and by objectifying some of the elements that go into reader judgment, the GCT has surely contributed to a solution of the problems of the essay test. Yet actually the Pearson report seems finally to conclude that: *no essay test is likely ever to be read with sufficient reliability nor is an essay a valid test of writing in the first place.*

It is unlikely that many teachers will concur in this conclusion. Yet it is an understandable one, and is surely a logical outcome once the testing of educational objectives fell into the hands of educational technicians. These technicians, originally seeking only to serve the classroom teachers, have by now come to dominate the classroom with the complexity and the importance of their functions. Their services to the teacher have been immense, yet applied techniques frequently lag behind the theory upon which they should be based and the mechanistic views of human psychology which prevailed during the thirties too often remain the basis of present-day practice. Writing is a creative process and the products of writing never have lent themselves readily to a mechanistic theory. They still do not, and when statistical analysis is applied to essay products without a regard for the difference between the results of an essay test and the results of an objective test, some confusion is likely to arise.

In recent years the tenor of educational psychology has changed considerably. The human organism is now regarded as a dynamic and creative whole rather than as simply the sum of a series of mechanistically defined parts. This shift in viewpoint has implications for language behavior. The mechanist concerns himself with the language *product* (and indeed, this has been and for the most part remains, the emphasis of the professional linguist). In keeping with the modern view of human psychology, it seems more feasible to concern oneself with the linguistic *process*. These observations are relevant to the grading of essays. It has been customary in the past to rank the essay products from Excellent to Nauseating without regard to what that product reveals about the writer. One of the departures in theory underlying the scoring of the GCT has been the insistence that the reader use the essay as a body of evidence for judging the writing process and hence that he attempt a judgment about the candidate's habitual writing behavior rather than attempting simply to rank the essays themselves. From this premise it followed logically that, since writing is a complex syndrome of human behaviors, the judgment of the essay should be broken down into separate qualities: Mechanics, Style, Organization, Reasoning, and Content, in the case of the GCT.

Reliability reviewed

It was felt that this kind of approach would lead to an increased objectivity of scoring and would result in a more meaningful report than would a single grade. The results of the two experimental readings of the GCT at Andover and Lawrenceville and of its first official reading at Haverford seem to bear out this assumption. If the judgments of readers are grouped so as to measure areas of agreement against areas of disagreement, the reader reliability coefficients at the three conferences turn out to be as follows:

Correlation of First Reader vs. Second Reader at Three GCT Reading Conferences

	M	S	O	R	C
Andover	.94	.88	.86	.90	.92
Lawrenceville	.85	.90	.87	.90	.93
Haverford*	.91	.92	.95	.83	.85

*Based on the sample of 370 cases randomly selected by Dr. Swineford in her analysis of the GCT reading; see Mr. Pearson's report, page 5, and "A question of reliability," page 12.

The reliability coefficients reported above do not reflect the influence of third readings, which were required whenever the first and second readers disagreed beyond one grade level. As Mr. Pearson has pointed out, the readers used border marks to indicate highness and lowness within a level as well as a mark to indicate that an essay was to be placed squarely within the grade level. If one were to look at reader reliability in another way, ig-

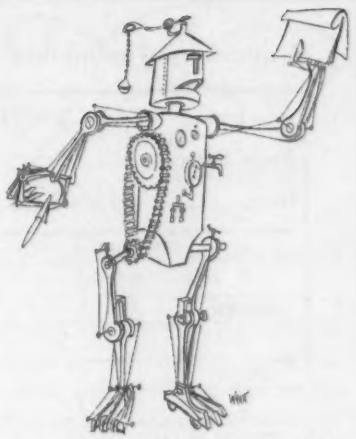
[1] Daniel Starch and Edward C. Elliott, "Reliability of the High-School Work in English," *School Review*, September, 1912, pp. 442-457.

noring the adjustments that were made when first and second reader data were grouped into areas of agreement and disagreement, but adding the influence of the third reader, thus reflecting in the reliability figures what actually went into the final grade, though basing these figures rigorously upon the "raw data," the following reliability coefficients result: Mechanics, .88; Style, .85; Organization, .82; Reasoning, .83; and Content, .83.[2]

The four and ten-point scale

Eighteen years ago John M. Stalnaker, who perhaps more than anyone else in the thirties worked for the essay test, and from whose pioneering efforts many of the features of the GCT were derived, came to the conclusion that if the scoring of an essay was to be considered satisfactory, "a reader reliability of around .90 is desirable."[3] The GCT has approached so closely the goal mentioned here by Mr. Stalnaker that we may probably consider the problem of reader reliability as solvable. It seems very likely that on the basis of the experience gained at Haverford and improved techniques of training, abetted by improved techniques of administration of the conference (to wit, the novel use of IBM equipment), reader reliability of the GCT can be improved up to a point. It should be noted, however, that the reading of essays is not precisely analogous to the comparison of samples of iron run through an electric furnace. Whereas in the latter case, correlations below .98 might be regarded as disastrous; in the former, correlations above .90 should be viewed with astonishment, and perhaps with suspicion.

One of the assumptions underlying the GCT is that in so complex an area readers cannot make extremely precise judgments, and it is for this reason that a four-point scale was adopted for



The human organism is not a series of mechanistically defined parts

each of the five qualities. The border indications allowed the readers have never been considered "real" grades, but rather indications of where the reader feels the student stands on a continuum of *good* to *bad*. GCT grades have never been reported except in terms of a four-point scale. An analysis of reader disagreements based upon a 10-point scale, as reported in Mr. Pearson's article, is extremely useful in assessing the actual work of individual readers at a reading conference, but may be misleading if used as a measure of the extent of agreement. I do not wish to enter into a discussion of what statistical technique is appropriate to various kinds of data. Possibly the simplest way to express the facts of the case is to point out that reader agreement at the two experimental conferences, before a third reading had been added, was about 80 per cent and at Haverford, where we were faced with the unforeseen situation of five tables of readers and about 4,300 essays (that is 8,600 readings) reader agreement dropped to about 76 per cent. One may treat these percentage distributions in various ways, and it is apparent that different results may be attained. The percentages remain constant.

The question of validity

An adequate essay test of writing is valid by definition; that is to say, it has face validity since it requires the candidate to perform the actual behavior which is being measured. It would be wrong, however, to assume that one essay test is as valid as any other since it is necessary, in order to achieve maximal validity in an essay, to secure a sample of the student's optimal writing behavior. For example, a test which poses an impossibly difficult problem will handicap, or even paralyze, many candidates; at the other extreme, a test which poses too simple or too concrete a problem may enmesh students, and sometimes the most able, in a simple-minded discussion which by no means reflects their true abilities. Considerable care was taken during the early days of the GCT to find means of eliciting an essay product which would be as valid as possible.

The readings which accompany the essay question were designed to accomplish two purposes: (1) to stimulate candidates to write, regardless of the level of their ability, and (2) to minimize among candidates differences in educational and personal experience. Moreover, we recognized that the way in which an essay is scored is relevant to its validity. Since we conceived that writing is a complex rather than a simple behavior, we felt that scoring the essays according to the five qualities would increase, not only the objectivity of reading, but also the validity of the reported scores. Further, the emphasis upon viewing each essay as evidence of a *process* rather than as a *product* to be ranked was a device for improving both the objectivity and the validity of the test.

Which teachers' estimates?

While there can be no doubt that the validity of the GCT may be improved, Mr. Pearson's conclusions about validity are surprising. These conclusions are based upon a study comparing GCT scores and teachers' estimates of writing ability. Similar investigations were conducted during each of the first two GCT experiments at Andover and Lawrenceville. The reported results of these earlier studies were not much different from the results Mr. Pearson now reports again. But the conclusions arrived at in this latter case are quite different from the conclusions which I drew from the earlier investigations.

The problem of validating an essay test is like pulling oneself up by one's own bootstraps, since the writing of an

[2] These figures are based upon a random sample of 200 cases from the Lawrenceville conference, since all of the raw data from the Haverford conference were not available to the writer. However, it seems reasonable to assume that the two conferences were not markedly different in results since the correlation coefficients in the above table are so similar.

[3] John M. Stalnaker, "Question VI — the Essay," *English Journal*, February, 1937, p. 134.

essay is the only direct measure of writing. When teachers' estimates of writing ability are used as a criterion against which to judge the GCT, we set up a situation like the following. In one case (the GCT), a group of highly qualified teachers of English meet together and, in training sessions, discuss and come to agreement about the meaning of the five qualities and of the four levels of scoring. They then proceed in a controlled situation to judge the essays. In the second case, a group of teachers of English, equally highly qualified, but working in separate schools, are asked to judge the writing competence of students, not in their usual way, but in the way it is done at the GCT reading conferences; that is to say, according to the five qualities and by the four levels, the meanings of which they may be expected to understand only in terms of their own experience and in terms of their own particular school situation.

It will not surprise many teachers that a low correspondence has been discovered consistently between GCT scores and teachers' estimates secured under these circumstances. I imagine, however, that the conclusion of Mr. Pearson that a carefully constructed essay test, painstakingly scored by competent teachers of English, is not valid will be surprising to many. The conclusion I have come to, after studying the results of these three similar investigations, is that the procedure of bringing English teachers together and allowing them to discuss what they mean by the objectives of English instruction is an extremely valuable one and, in a short time, leads to a remarkable consensus.

The sum of the parts

However, Mr. Pearson's conclusions with respect to validity add one surprise upon another. For example, we find him saying that the essays written in response to the GCT may test ability in Mechanics and Style, but do not test ability in Organization, Reasoning, and Content. As a matter of fact, we have not been able to produce a test which will allow the student to write an essay in which he exhibits his competence in mechanics and style, but organizes nothing, makes points without reasoning, and produces about a thousand words devoid of content. Actually, the Pearson report is based upon averaged coefficients secured from 11 different schools. In the Andover experiment I found about the same results as Mr. Pearson reports when I averaged teachers' estimates (not from 11, but from 29 secondary schools). The differences among the schools were tremendous, ranging, for

A question of reliability

GRADING SHEET

No. of Paper_____ No. of Reader_____

Date_____ Time Begun_____ Time Finished_____

	Mechanics	Style	Organization	Reasoning	Content
4	X				
BORDER	∧				
3					
BORDER					
2					
BORDER					
1					

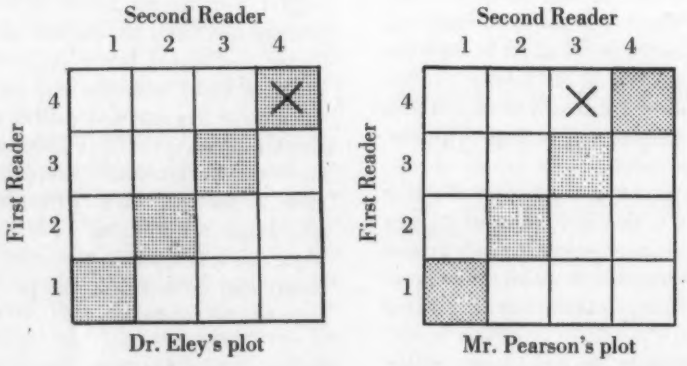
Place papers in categories 1 to 4 with an X; border may be marked *only* ∧ or V.

GCT grading sheet illustrates case described below in which two readers assign different scores in Mechanics category. In practice, each reader uses a separate grading sheet and indicates scores in all five categories

The startling difference between the GCT reader reliability correlations reported in their articles by Mr. Pearson (page 5) and Dr. Eley (page 10) arises, among other things, from a difference between the authors as to what constitutes "agreement" between readers.

In the oversimplified illustration one reader scores the essay "4" in Mechanics and a second reader indicates by the carat (∧) in the borderline space that he considers it a "high 3" in Mechanics. Dr. Eley accepts this as agreement and *before* computing his correlation for Mechanics, plots it as an instance of reader agreement, as indicated by the cross mark in the shaded area below. Mr. Pearson considers the "4" and "high 3" scores to be a case of reader disagreement and plots his mark in the unshaded area to indicate this.

Since the correlations obtained are essentially numerical expressions of agreement, the correlations reported by Dr. Eley are higher than those of Mr. Pearson.



Dr. Eley's plot Mr. Pearson's plot

example in Mechanics, from .16 to .82. If, instead of averages, I had reported the correlation between GCT and teachers' estimates, selecting for each quality the best school, the correlations would have been: Mechanics, .82; Style, .67; Organization, .72; Reasoning, .60; and Content, .58. The lowest correlations I could have reported would have been: Mechanics, .16; Style, .41; Organization, .22; Reasoning, .42; and Content, .33.

There is considerable evidence that the GCT scores in Organization, Reasoning, and Content have some independence. For example, in the Andover experiment it was found that the score in Reasoning correlated with the College Board Social Studies Achievement Test at .31, while the score in Content correlated at .68. On the other hand, the GCT score in Reasoning correlated with the College Board Mathematics Achievement Test at .65, whereas the score in Content correlated at only .24. Further, in all three conferences the intercorrelations among qualities have demonstrated reasonable independence, the highest ordinarily being between Reasoning and Content at about .75.

Language as a process

Research in language behavior historically has been centered upon a concern with language products, and this generalization applies not only to the teacher and the maker of tests, but to the professional linguist, the sociologist, the psychologist, the anthropologist, and even to the philosopher.

Earle G. Eley is examiner in English and Humanities at the University of Chicago and is currently teaching English composition there. Dr. Eley led the experimental work on the General Composition Test and has been its principal interpreter and spokesman. He was chief reader of the first regular administration of the test last May.

We can evaluate language behavior only in terms of the total process

Probably, of all of these the teacher of writing has come closest to dealing with the process of language, though he has rarely, if ever, reported the procedures of his classroom in theoretical terms. This situation is a logical consequence of the recent history of language studies. During the twenties, thirties, and early forties, when the prevailing psychology in the United States was mechanistic, that is to say, behavioristic, the professional linguist was busily engaged in the attempt to make an exact science from a humane study, and was happy enough to exclude the investigation of a psychological basis for language behavior from his enterprise. It is not surprising, then, that whatever statement about the psychological basis of human behavior we find among the linguists of this period, turns out to be a direct and naive reflection of stimulus-response psychology.[4]

Thus, we have the science of linguistics depending upon another science, psychology, for its theoretical basis. The test maker who engaged in constructing tests of writing ability was doubly handicapped in that, nurtured by behavioristic psychology, he found that the professional linguist afforded him no more than a reflection of the same climate. An applied scientist, he had little to apply. That he should have failed to recognize that language behavior is actually an integral part of the total effort of each personality to realize itself and is, even at its lowest levels, more closely related to creative endeavor than to conditioned activity, is pardonable — but an error.

Few modern psychologists would accept the label mechanistic, or behavioristic. The notion that some kinds of human behavior are complex syndromes of activity which represent wholes that are more than the sum of their parts is today commonly accepted. Language clearly is such a human activity. When we investigate or attempt to evaluate language behavior, we can do it only in terms of the total process. Were it possible to reduce these complex and mysterious aspects of the human personality to elements capable of exact measurement, our universe would be a simple, secure and some may think, a happy one. Too often, we have achieved precision in the measurement of human behavior by mechanizing the elements of that behavior. Such a procedure is effective only part of the time: the creative or dynamic aspects of the human personality resist atomization. They must be assessed as wholes or not at all.

Objectives of the test

Thus, it seems unlikely that objectively scored tests can be developed to measure validly the ability to write, which is a creative aspect of the human personality. The GCT attempts to measure this ability as a totality. We sought to develop as much precision as possible, but we knew from the beginning that writing cannot be measured with as much precision as can, for example, the reaction time of knee jerks. Moreover, we have avoided deluding ourselves into mechanizing the testing or reading procedures — a device of most objective measures of creative activity — in order to produce the happy security of spuriously high statistics.

The GCT ought to be judged, I should think, in terms of whether (1) it satisfies a need in American testing, (2) it satisfies a demand expressed by teachers of English, and (3) it achieves a degree of objectivity appropriate to the kind of instrument that it is. There is no doubt that from some points of view, we could find reasons for abandoning forever the essay test. Unfortunately, the corollary to this action is the decline in the teaching of the art of writing. This is a serious consequence both for the individual and for our society. The ability to use language creatively is one of the last strongholds of the individual against a growing tendency toward conformity.

[4] See, for example, Leonard Bloomfield, *Language*. (New York: Henry Holt and Co., 1933), pp. 31-32.

BY MARY E. CHASE

School visiting made easier—and better

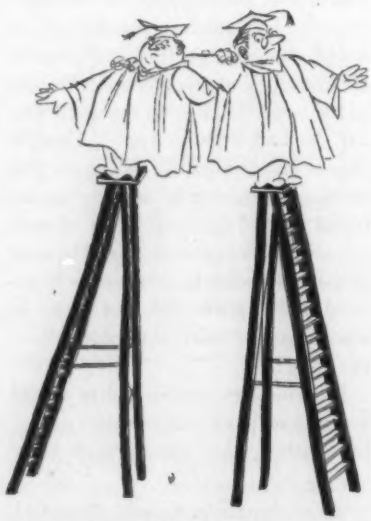
By working together college representatives can reach more schools, inform more students, eliminate duplication

"And now girls, Miss Wellesley will talk to you about liberal arts colleges for women." This introduction, made several years ago by the principal of an eminently respectable school, unnerved me considerably, for I had always pictured Miss Anyplace as a curvaceous tidbit, scantily clad in the barest-of-essential bathing suits, who walked rather than talked, and I was obviously miscast. It later developed that the principal in question was so completely startled when, in accepting her invitation to speak to the women students, I had suggested that I would prefer to talk about other colleges as well as about the one which employed me, that she forgot my name and remembered only the college.

I doubt that this could happen today, for there is a general acceptance among representatives of women's colleges and a growing recognition on the part of school people that what is good for one college is usually good for another, and that a high degree of cooperation with one another is not only possible but extremely desirable.

School visiting by college representatives still has room for vast improvement and refinement. We shall probably have always with us, like the poor, such extremes as what is described by an acquaintance of mine as "the occasional foray made into the field by last year's football hero," the dashing Beau Brummell in his Buick who charms the family and paints enchanting pictures of "living happily ever after" to the little girl whose social maturity has been slow and painful, and the erudite physics professor with baited hook who has been torn from his atomic pile and thrown into an unfamiliar world by the president of his college, charged with the responsibility of increasing the number of majors in physics at Siwashette.

Even the normal (if there is one) college representative with beaten-up suitcase and bulging brief case, weary of trains, dismal hotel rooms and occasional orgies of creamed chicken and peas (if sufficient alumni can be prevailed upon to gather to meet him) meets some extremes in the schools. There is the stony glare of the grim headmistress who plants herself firmly in the front hall (if the college visitor has succeeded in getting that far) and says, "all our girls *have* plans," but there is also the slap-happy principal who upon the arrival of the college visitor dismisses classes and asks the college visitor to address the junior high school and senior high school boys and girls on "Education." Some school people are buried in an unending mass of administrative detail and can quote the exact number of square inches of floor space in the gymnasium and the number of paper towels used in the boys' washroom per week, but throw up their hands in despair if one asks for information about their graduates in college or the content of the senior English course. Most of us in the business have seen them all at one time or another.



A high degree of cooperation is desirable

Reasons for visiting

The purpose of the school visit is, of course, twofold—give and take. The visitor secures information about the school which helps him to interpret the records of future candidates for his college trained in the school, and provided he is skillful, the college visitor gives the high school principal, counselor or dean, an up-to-date picture of his institution and an understanding of what is considered a well-qualified candidate. The school visit by the college representative provides an excellent opportunity for a face-to-face discussion of individual candidates.

While an optimum arrangement from one point of view would be an annual visit to each school, it is obviously impossible for most colleges to support a sufficiently large professional staff to make this possible. (School principals already harrassed by flocks of college visitors take comfort! Most colleges cannot afford to expand their admission staffs.)

The compromise arrangement adopted by a number of colleges is to set up a long-term travel plan. The college representative works out a schedule of travel to a different section of the country each year, so that the major portion of the country from which the college draws its students can be visited once in every three to five years. The long term travel is sup-

plemented by annual visits to individual schools or areas which present a substantial number of candidates each year and which might be referred to as "feeder areas" or "feeder schools."

With the vastly increased numbers of applications anticipated in the next few years and with applications which will come in many cases from new schools, the necessity for a continuation and extension of school visits is obviously essential. Yet at the same time, if colleges are to maintain the quality of their academic programs, they are faced with the continuing need for economies and cannot in most instances afford to increase their admission staff proportionally. Is it presumptuous to suggest that a college desires to be represented by a well-informed and well-trained person—preferably one who has had previous teaching experience in either the secondary or collegiate field or both, and that such individuals are not easy to find nor are they inexpensive?

Each college which operates a consistent program of school visiting has in its possession a considerable amount of information about the various secondary schools which have been visited. The gathering of this information by colleges across the country is very expensive, especially when one considers that a single college visitor could easily make copies of his information about a given school or schools and send it to five colleges, let us say, at considerably less effort, energy, and expense than is currently required when five college visitors secure the same information separately for their records.

Perhaps the emergency presented by these two phenomena, increased demands on the admissions officers and duplication of information and effort, points as often only an emergency can, to a very simple and considerably more efficient method of handling this whole problem.

A cooperative effort

Is it too much to expect Miss Wellesley, Mr. Kenyon, or Mr. and Mrs. Pomona to be informed about other colleges of their type so that they can represent more than their own particular institution? Surely no one person could be expected to have complete factual information on *all* colleges, but there is no reason to believe that a professional in one college cannot represent similar sister or brother institutions. I am not proposing, for example, a College Board Representative to represent College Board member colleges, but rather the formation of small informal *groups* of colleges which have common programs and interests and the desire to cooperate with one another in the presentation of their programs to schools and to candidates in the schools.

It is important that such college groups be kept small, since too large and too diverse a group will rob each institution of its individuality. Unless the group is small and reasonably homogeneous, it is quite impossible for a single representative of any one of the colleges to be adequately informed, and his usefulness is thereby reduced. It is important that such a grouping be kept informal, for too much organization can defeat the very purpose of the cooperative venture, and the values to be achieved can be smothered by endless red tape and a sheeplike conformity. Each cooperating college in the group must maintain its own integrity and be allowed "to be itself."

How would such a plan work? When several colleges recognize themselves as a natural group, it would be comparatively simple to determine in what area, or areas, of the country school visits are most essential, and by cooperative means a master travel plan could be worked out so that not more than one representative would visit the schools in a particular area during a given academic year. If, for example, there were three colleges in the group, the coverage would be tripled. It would be a comparatively simple matter for the group to work out jointly in advance some agreement as to the type of information desired about each school, so that all members of the group could share it. It would be equally important for the colleges to work out together, in advance, policies with respect to answers to questions regarding unusual programs of sec-

The purpose of the school visit is twofold

Mary E. Chase, executive vice president and director of admission of Wellesley College, is particularly interested in the college's relations with prospective students, schools, and alumnae. She has held school and college posts and served in the WAVES during World War II before going to Wellesley, in 1946, as director of admission.

ondary school study, exceptions to requirements, and specific problems relating to each individual college. The important similarities of the colleges should be familiar to all members of the group, as should important dissimilarities. One should be able to present *the colleges* directly to counselors and students, referring specific questions to individual catalogues or individual offices.

While the actual geographical location of the colleges in any given group would be much less important than the internal similarities of the colleges in the group, there would be certain advantages to be gained if the colleges were located in different parts of the country. An ideal grouping might be a southern college, a New England college, a midwestern college and a far western college. There would be obvious advantages to each college in the group in acquainting students from widely diverse geographical areas about the programs in each college.

The time is now

There are those who would argue that since the colleges have a seller's market, why visit at all? Why recruit students when application lists are already bursting the files? The reason is that school visits are not recruiting ventures, but rather attempts to improve understanding and communication between schools and colleges. From the college's point of view, school visiting is most important when the supply of qualified candidates exceeds the number of places available to them, for it is at this time that rumors, misunderstanding, and folklore are at their height, and the time when it is essential that there be close understanding between the schools and colleges and open lines of communication.

It is hoped that, in the future, school systems will recognize and adopt programs similar to the one currently in operation in Atlanta, Georgia, where a forward-looking administration has appointed an experienced school man as a liaison person between the Board of Education and the colleges attended by Atlanta students. This liaison officer is the authorized representative not of one school, but of all the public high schools in the county, and has close

Too much organization can defeat the very purpose of the cooperative venture

association with students already in college and with the principals and guidance officers in each school. Advance notice to the liaison officer of the arrival in Atlanta of a representative of one of the "groups" described above would facilitate an economical use of the time of both the college representative and guidance personnel, since a single, rather than separate meetings of several representatives, could be easily arranged and many problems and questions could be aired in a relatively short time to the advantage of the already overburdened counselors.

It is conceivable that a cooperative effort among the admissions representatives might result in a similar effort among their alumni to search out and acquaint talented young people with the opportunities available to them. In spite of swollen application lists, probably no college ever has too many *good* candidates. What better source of information is there about these good candidates, outside of the school authorities themselves, than educated members of the community? Alumni who are loyal to their college and wish to "do something" for it could be invited to participate in a program which does not ask them for money. Many young alumni just starting out in business and professions and unable to assist their college financially might contribute generously to such a cooperative effort. The colleges would benefit by their participation both directly in terms of students, and indirectly in terms of interest.

Cooperate!

Recruiting is a naughty word.
So's competition I have heard.
Let's group ourselves across the nation,
And try, for once, cooperation.

Additional benefits

Such a grouping of colleges might have other very valuable side results. One of these might be the lifting of the veil of secrecy and the elimination of the atmosphere of hocus-pocus which quite unintentionally surrounds the admissions process in many colleges. In order to discuss the admissions *policies* of the colleges in a given group, one must know what their policies are—which might encourage some admissions committees to collect their thoughts and express them on paper. Differences in *procedure* in individual colleges in the group would come to light, and one college might profit from the experiences of another. The whole area of requirements, too, would come under scrutiny and enable the institutions in the group to pool their studies and have a better basis upon which to determine the actual courses essential for admission.

This proposal would seem to deny the maxim, "Never urge people to do together what the self-reliant among them can do better alone" but in the present and future emergency in colleges, each college in such a group could profit from the union and could at the same time be allowed to retain its individuality.

There are dangers, of course, but not the same dangers which business or industry would encounter in such a grouping. Certainly each college in such a group must be constantly on guard against mass anonymity and excessive standardization. But as long as faculty members of colleges act as individuals and insist that their colleges "be themselves" the danger of encroachment by the monster of Big Education is not great.

The plan certainly seems worth trying.

How to "get by" written examinations

BY STEPHEN LEACOCK

Sir: You are kind enough to refer to certain of my writings in regard to the difficulties and fallacies of written examinations. You ask me if there is any way—if I have your phrase right—to "get by." I think there is.

Every student should train himself to be like the conjurer Houdini. Tie him as you would, lock him in as you might, he got loose. A student should acquire this looseness.

For the *rudiments* of education, there is no way round. The multiplication table has got to be learned. They say Abraham Lincoln knew it all. So, too, the parts of speech must be committed to memory, and left there. The

The multiplication table has to be learned

names of the Wessex Kings from Alfred (better Aelfoydd) to his Danish successor Half-Knut should be learned and carefully distinguished from the branches of the Amazon.

But these rudiments once passed, education gets easier and easier as it goes on. When one reaches the stage of being what is called a ripe scholar, it is so easy as to verge on imbecility.

Now for college examinations, once the student is let into college, there are a great number of methods of evasion. Much can always be done by sheer illegibility of handwriting and by smearing ink all over the exam paper and then crumpling it up into a ball.

But apart from this, each academic subject can be fought on its own ground. Let me give one or two examples.

Here, first, is the case of Latin translation—the list of extracts from Caesar, Cicero, etc., the origin of each always indicated by having the word Caesar, etc., under it. On this we seize as our opportunity. The student doesn't need to know one word of Latin. He learns by heart a piece of translated Caesar, selecting a *typical* extract, and he writes that down. The examiner merely sees a faultless piece of translation and notices nothing—or at least thinks that the candidate was given the wrong extract. He lets him pass.

Here is the piece of Caesar as required:

These things being thus this way, Caesar although not yet did he not know neither the copiousness of the enemy nor whether they had frumentum, having sent on Labienus with an impediment he himself on the first day before the third day, ambassadors having been sent to Vercingetorix, lest who might which, all having been done, set out.

Cicero also is easily distinguished by the cold, biting logic of his invective. Try this:

How now which, what, oh Catiline, infected, infracted, disducted, shall you still perfrage us? To what expunction shall we not subject you? To what bonds, to what vinculation, to how great a hyphen? I speak. Does he? No.

Cicero. In (and through) Catiline

The summation of what is called the liberal arts course is reached with such subjects as political theory, philosophy, etc. Here the air is rarer and clearer and vision easy. There is no trouble at all in circling around the examiner at will. The best device is found in the use of quotations from learned authors of whom he has perhaps—indeed, very likely—never heard, and the use of languages which he either doesn't know or can't read in blurred writing. We take for granted that the examiner is a conceited, pedantic man, as they all are—and is in a hurry to finish his work and get back to a saloon.

Now let me illustrate.

Here is a question from the last Princeton examination in Modern Philosophy. I think I have it correct or nearly so.

"Discuss Descartes' proposition, 'Cogito ergo sum,' as a valid basis of epistemology."

Answer:

"Something of the apparent originality of Descartes' dictum, 'cogito ergo sum,' disappears when we recall that long before him Globulus has written 'Testudo ergo crepito,' and the great Arab scholar, Alhelallover, writing about 200 Fahrenheit, has said, 'Indigo ergo gum.' But we have only to turn to Descartes' own brilliant contemporary, the Abbé Pâté de Foie Gras, to find him writing, 'Dimanche, lundi, mardi, mercredi, jeudi, vendredi, samedi,' which means as much, or more, than Descartes' assertion. It is quite likely that the Abbé was himself acquainted with the words of Pretzel, Wiener Schnitzel and Schmierkase; even more likely still he knew the treatise of the low German, Fisch von Gestern, who had already set together a definite system or scheme. He writes: 'Wo ist mein Bruder? Er ist in dem Hause. Habe ich den Vogel gesehen? Dies ist ein gutes Messer. Holen Sie Karl und Fritz und wir

Fight each subject on its own ground

werden alle ins Theater gehen. Danke Bestens.' "

There, one can see how easy it is. I know it from my own experience. I remember in my fourth year in Toronto (1891) going into the exam room and picking up a paper which I carelessly took for English philology; I wrote on it, passed on it and was pleasantly surprised two weeks later when they gave me a degree in Ethnology. I had answered the wrong paper. This story, oddly enough, is true.

Reprinted from the "To the Editor" columns of *The Daily Princetonian*, Wednesday, January 26, 1938.

BY JOHN R. VALLEY

The case against freshman flunk-outs

How many applications for admission are needed to enable a college to select a class of certain size which meets a certain standard of scholastic performance?

While this question cannot be answered precisely and to the man, it was found at Case Institute of Technology that predictions can be made which are very useful administratively. A good understanding of the way a college's selection criteria actually operate may show, for example, that only 500 rather than 750 applications need be evaluated to obtain the desired class—and keep it. A difference of this kind may also be considered somewhat more than statistically significant when translated in terms of the size of an admissions office staff and its work load.

The basic choice the college has is whether to establish its admissions goals in expectation of losing a certain percentage of students through failure, or to employ predictive procedures which identify those most likely to succeed academically. In the first case the emphasis is on numbers and the goal is a freshman class large enough to include that certain percentage of students whose departure later will be accepted as natural and normal and characterized as "attrition." The more selective method, however, not only locates those individuals whose survival in college is likely, but finds more of them in any given number of applicants.

The Case Institute experience suggests that although attrition may be a necessary evil, selective measures enable a college to plan its recruitment and admissions activity much more wisely. An institution can know with very little doubt how many applications will be needed and how much effort will be required to get them in order to produce a given number of acceptable applicants. It can also know how the attrition rate would vary if the admissions standards were changed to produce a larger or smaller entering class.

Last year at Case, for example, 25% of the freshman class were separated for reason of poor scholastic performance or withdrew voluntarily, most of them also for academic reasons. How would selection by an appropriate battery of College Board tests, or by a group of tests which we can identify for the present as the Case Battery, or by high school rank in class, influence that attrition rate? Let us consider the accompanying table which shows the rate of attrition expressed as a cumulative percentage of the freshman class grouped by deciles according to grade point averages predicted on the basis of the three criteria mentioned above.[1]

For purposes of illustration, assume that Case wished to cut the total freshman attrition rate in half. Assume also that the 370 students who made up the freshman class constituted instead a group of applicants from which we were ultimately to admit a class. The table shows that an attrition rate of 12% existed in the first *six* deciles of the class when it was ranked according to point averages predicted by the College Board battery. The same attrition, 12%, is found in the first *five* deciles when the class was ranked according to point averages predicted by the Case Battery. The 12% is found in the first *four* deciles when the ranking was by point averages predicted by high school rank in class.

This means that the College Board tests would have selected a class of 222 students out of the 370 applicants, the rate of attrition being 12%. The Case Battery would have selected 185, and the rank in class criterion 148 students, for the same attrition.

Thus, selection by means of the Board tests would have given Case the opportunity to admit 74 more students into a class with a 12% attrition than could have been secured by means of high school rank in class and 37 more students than would have been found by the Case Battery.

The advantage here lies in the capacity of the Board tests to identify, within a particular group tested, a larger number of individuals who can meet a certain performance standard. Taking this advantage, Case would have fewer students to recruit, fewer applications to process, fewer admissions interviews, fewer high school transcripts to study—fewer of everything except qualified students who would complete their freshman year satisfactorily.

In order to actually net a class of 450 freshmen which would have an attrition only one-half that of a year ago, it is estimated that Case would need twice as many completed applications with Board test results as it had

[1]The statistical analyses for this study were done by Dr. Fred C. Leone, assistant professor of mathematics, Case Institute of Technology.

John R. Valley was assistant dean of students at Case Institute of Technology at the time the study described above was conducted. He left Case last December to become assistant project director for the College Board at the Educational Testing Service.

in September 1953. Two and one-half times as many processed with the Case Battery and three times as many from which selection would be made by high school rank in class would be needed to yield the 450 freshmen.

The data on which this illustration is based are from the latest study to reflect Case's long-standing interest in the prediction of scholastic success, which dates back to the early twenties. Following World War II, the Institute used the Pre-Engineering Inventory as a selection device for admission. With the elimination of nation-wide testing facilities for the Pre-Engineering Inventory, we began to experiment with the Pre-Engineering Ability Test.

Looking forward to the probable figures of the college-going population of the next decade, we felt that the need for better prediction would increase in the future. We were also aware of the time lag between the initiation of a research program, the attainment of meaningful results, and the application of useful research findings. It was obvious that any prediction study which depended on first year academic performance for its criterion of success would take at least a year before that part of the data was available.

In 1953, with the cooperation of the College Board and the technical assistance of the Educational Testing Service, we began a study designed to show the comparative use of the predictors whose application to a particular admissions situation was described above. The Board provided a special campus administration of the Scholastic Aptitude Test, which unlike the other tests used in the experiment is not available in the Board's placement test program.[2] The ETS staff contributed importantly to the success of the plan with its suggestions for the design and methodology of the research.

The predictors of scholastic success selected for study were:

1. High school record, as measured by rank in school graduation class
2. College Board tests (Scholastic Aptitude Test and the Achievement Tests of Chemistry, Physics, and Advanced Mathematics)
3. The "Case Battery" which had been used at Case largely for post-admission guidance and consisted of:
   a. Pre-Engineering Ability Test
   b. Cooperative General High School Mathematics Test, Revised Series Form O
   c. Cooperative Physics Test, Revised Series Form Y
   d. Cooperative Chemistry Test, Revised Series Form Y
4. Lapse of time since completion of high school physics
5. Lapse of time since completion of high school chemistry

[2]The College Board customarily supplies testing materials and scoring and reporting services at no cost to those colleges which wish to explore the possibility of using the tests as an admissions requirement and agree to conduct formal validity and norms studies at their own expense and report the results to the Board.

Rate of attrition for the freshman year (expressed as a cumulative per cent)*

Number of students (37 in each decile)	Deciles of the class according to predicted point average	Basis of predicted point average: College Board battery plus high school rank and recency variables	Basis of predicted point average: Case battery plus rank in class and recency variables	Basis of predicted point average: High school rank in class
37	1	3%	5%	7%
74	1 and 2	5	5	8
111	1 through 3	8	11	9
148	1 through 4	8	11	12
185	1 through 5	9	12	14
222	1 through 6	12	16	17
259	1 through 7	17	18	21
296	1 through 8	19	20	22
333	1 through 9	22	23	23
370	1 through 10	25	25	25

* *Attrition = separations and withdrawals for all reasons.*

The entire class of 370 students who enrolled in September 1953 took the College Board tests during their orientation week at the start of the first term. The Case Battery data were available for the whole group as follows: 22% tested with the full battery prior to admission; 37% took the three achievement tests prior to admission and the Pre-Engineering Ability Test during orientation week; 41% tested with the full battery during orientation week. In other words, 41% of the class was admitted exclusively on the basis of their secondary school records. Either complete or partial test data from the Case Battery for the remaining 59% were available at the time they were admitted.

In the study, we compared each of the principal predictors with scholastic success, which was measured in terms of the freshman year grade point averages. The correlations for each predictor were as follows:

Predictor	*Grade correlation*
College Board tests	
SAT-Verbal	.331
SAT-Mathematical	.442
Chemistry	.438
Physics	.408
Advanced Mathematics	.401
Case Battery	
Pre-Engineering Ability	.341
Cooperative Mathematics	.356
Cooperative Chemistry	.349
Cooperative Physics	.298
High school data	
Rank in graduating class	.299
Recency of physics study	.105
Recency of chemisty study	.089

This showed that the College Board tests did the best job of predicting and that the Case Battery ran just a little better than rank in class. The recency variables were revealed to have low correlations with first year grades.

Using this information, as well as intercorrelations between the predictors, we found that a multiple correlation of .600 could be achieved by combining the Board test and high school data. A multiple correlation of .517

resulted from a combination of Case Battery and high school data. It was apparent that our prediction of academic success could be improved.

The computation of correlation coefficients also enabled us to derive equations for the prediction of first year grade point averages. Three such equations were derived and three predicted point averages calculated for each student on the following bases:

1. College Board test scores, together with high school rank in class and recency of completion of study of physics and chemistry
2. Case Battery test scores, together with high school rank in class and recency of completion of study of physics and chemistry
3. High school rank in class alone

The freshman class was divided into deciles, as illustrated by the accompanying table, on the basis of each of the predicted averages. Students in the top 10% of the class on the basis of their predicted averages were placed in Decile 1, those in the next highest 10% in Decile 2, etc. The table shows the number of chances out of 100, computed from observed grade point averages, of a student being in good academic standing both semesters when his performance is predicted by each of the three point averages.

Predictions based on College Board scores are shown by the table to be more definitive than those which depend on the other criteria of selection. If a student fell in the first decile according to his point average calculated by the College Board equation, for example, he had 97 chances in 100 of being successful. His chances with the Case Battery were 92 out of 100, and with the high school rank in class 88 out of 100. The difference in degree of definition becomes especially noteworthy at lower points on the table. The student who fell in Decile 10 according to the College Board predictor, for example, had only 23 chances out of 100, while the Case Battery shows 40 and high school rank alone 43 chances.

The college which wants to eliminate unpromising candidates before they become failing students and identify those most likely to succeed academically in any given number of applicants will find this kind of appraisal of its admissions criteria most rewarding—much more rewarding than the cold bewildered stare of a faculty member who is informed merely that a correlation of .45 has been found to exist between SAT-Mathematical and first term grade point averages.

Prediction of First Year Scholastic Success at Case Institute of Technology

If ranked in	According to point average predicted by	Chances in 100 That Student Would Be in Good Standing Scholastically Both Semesters
		0 10 20 30 40 50 60 70 80 90 100
Decile 1	———	
	- - - - -	
	•••••••	
Decile 2	———	
	- - - - -	
	•••••••	
Decile 3	———	
	- - - - -	
	•••••••	
Decile 4	———	
	- - - - -	
	•••••••	
Decile 5	———	
	- - - - -	
	•••••••	
Decile 6	———	
	- - - - -	
	•••••••	
Decile 7	———	
	- - - - -	
	•••••••	
Decile 8	———	
	- - - - -	
	•••••••	
Decile 9	———	
	- - - - -	
	•••••••	
Decile 10	———	
	- - - - -	
	•••••••	

——— = *CEEB Battery, High School Rank in Class and Recency of Completion of High School Physics and Chemistry.*
- - - - - = *Case Battery, High School Rank in Class and Recency of Completion of High School Physics and Chemistry.*
••••••• = *High School Rank in Class alone.*

College Board member colleges

with special reference to participants in the College Scholarship Service and subscribers to the Candidates Reply Date

Of the 162 member colleges of the College Board listed below, 95 are participating in the College Scholarship Service. These institutions are designated by a check mark (✓).

Member colleges subscribing to the Candidates Reply Date agreement are indicated by a dot (•). These 120 institutions have agreed not to require any candidate admitted as a freshman to give notice before May 18, 1955 of his decision to attend one of the subscribing colleges or to accept financial aid from it. Colleges listed with an asterisk (*) except scholarship applicants from the Candidates Reply Date.

The Candidates Reply Date, originally known as the Uniform Acceptance Date, was first used in 1948 when eight colleges signed the agreement for that year. It was inaugurated in order to give candidates ample time to consider all opportunities available to them. Now, as in earlier years, however, the candidate is encouraged to notify all institutions to which he has applied as soon as he definitely decides which college he wishes to attend.

•✓Adelphi College
Agnes Scott College
•✓Albertus Magnus College
•✓Alfred University
•✓Allegheny College
✓Amherst College
•✓Antioch College
•✓Bard College
•✓Barnard College
✓Bates College
•✓Beaver College
Beloit College
• Bennington College
Boston College
• Boston University
✓Bowdoin College
✓Brandeis University
•✓Brown University
•✓Bryn Mawr College
•✓Bucknell University
• Caldwell College
•✓California Institute of Technology
•✓Carleton College
✓Carnegie Institute of Technology
• Catholic University of America
•✓Cedar Crest College
• Chestnut Hill College
•✓Claremont Men's College
•✓Clark University
•✓Clarkson College of Technology
•✓Colby College
✓Colgate University
• College of Mount Saint Vincent*
• College of New Rochelle*
• College of Notre Dame of Maryland*
• College of Saint Elizabeth*
• College of William and Mary
College of Wooster
•✓Columbia College
•✓Connecticut College
Cooper Union for Advancement of Science and Art
•✓Cornell University
•✓Dartmouth College
✓Davidson College
•✓Denison University
• DePauw University
•✓Dickinson College
• Drew University*
• Duke University*
•✓Dunbarton College of Holy Cross*
• Elmira College*
•✓Emmanuel College
Emory University
•✓Fordham College
✓Franklin and Marshall College
Georgetown University
George Washington University
• Georgian Court College*
• Gettysburg College
•✓Goucher College
Grinnell College
✓Hamilton College
•✓Harvard College
•✓Haverford College
•✓Hobart and William Smith Colleges
•✓Hollins College
• Hood College
• Immaculata College
•✓Jackson College for Women
Kalamazoo College
✓Kenyon College
• Knox College*
✓Lafayette College
• Lake Forest College
✓Lehigh University
• Lewis and Clark College
•✓Manhattan College
•✓Manhattanville College of the Sacred Heart
• Mary Baldwin College
• Marymount College*
• Marywood College*
•✓Massachusetts Institute of Technology
• McGill University
Michigan State College
✓Middlebury College
•✓Mills College
•✓Mount Holyoke College
✓Muhlenberg College
Newark College of Engineering
• Newcomb College
•✓New Jersey College for Women
New York University
•✓Northwestern University
✓Oberlin College
•✓Occidental College
✓Ohio Wesleyan University
•✓Pembroke College in Brown University
•✓Pennsylvania College for Women*
Pennsylvania State University
•✓Pomona College
•✓Princeton University
• Providence College*
•✓Radcliffe College
•✓Randolph-Macon Woman's College
•✓Reed College
•✓Regis College
•✓Rensselaer Polytechnic Institute
✓Rollins College
• Rosemont College*
• Russell Sage College
•✓Rutgers University
• Saint Joseph College (Connecticut)
• Saint Joseph College (Maryland)
• St. Joseph's College for Women
✓St. Lawrence University
• Saint Mary's College
• Salem College
•✓Scripps College
• Seton Hill College
•✓Simmons College
•✓Skidmore College
•✓Smith College
•✓Stanford University
• Stevens Institute of Technology
•✓Swarthmore College
•✓Sweet Briar College
• Syracuse University
✓Trinity College (Connecticut)
• Trinity College (Washington, D. C.)
•✓Tufts College
✓Union College
University of California
• University of Chicago
• University of Colorado
University of Connecticut
University of Denver
✓University of Massachusetts
• University of Michigan
University of Notre Dame
•✓University of Pennsylvania
•✓University of Redlands
•✓University of Rochester
•✓University of Southern California
• University of the South*
University of Vermont and State Agricultural College
• University of Virginia
• Ursinus College
•✓Vassar College
• Villanova University*
•✓Wagner College
• Washington and Jefferson College*
• Washington and Lee University
•✓Wellesley College
•✓Wells College
✓Wesleyan University
• Western Reserve University
•✓Wheaton College
•✓Whitman College
•✓Whittier College
✓Williams College
•✓Wilson College
•✓Yale University

NEWS OF THE COLLEGE SCHOLARSHIP SERVICE

Early activities

First reports: By mid-February, the parents of 10,000 scholarship applicants had sent their College Scholarship Service financial statements to the Service's offices in Princeton and Los Angeles. There photographic copies were made and sent to those colleges named by the parents on the statement.

A survey of the 8,000 statements received by the end of January showed that the average number of colleges listed to receive copies was 2.39 in the Princeton service area and 2.41 in Los Angeles. These figures correspond closely to the average number of colleges listed by the entire College Board candidate group to receive test score reports. The scholarship-candidate ratio was expected to increase only slightly if at all as deadlines for financial aid applications drew closer.

Before copies of the statements are sent to colleges they are scanned for omissions and apparent inaccuracies and these are indicated by the Service on the form. The parents are requested to clarify such items and their explanations are transmitted to the colleges. The first scanning experience by a small staff working on a part-time basis showed that 10 statements an hour could be inspected. The record was 112 forms for one person during an eight-hour day. About one-quarter of the cases required follow-up correspondence with the parents in order to complete the information requested or eliminate ambiguities. The parents' responses to these requests were made with good will and care.

Information available: A preliminary version of a CSS manual which outlines procedures which may be used in estimating the financial need of students has been distributed to colleges participating in the Service. Other activities have included meetings by several groups of college scholarship and admissions officers to discuss computational methods and the scheduling of a general meeting by the Service on April 5 at which college representatives may report their experiences and present their comments and criticisms of the Service.

The CSS welcomes requests from colleges for information concerning the use of the financial statement. Inquiries should be addressed to the College Scholarship Service, Box 176, Princeton, New Jersey, or Box 27896, Los Angeles, California. A general description of the Service's purpose and activities is also available in a leaflet which will be sent to school and college officers on request.

List of applicants: The Service has been requested by several small groups of colleges to prepare lists of their joint scholarship applicants. Lists of

this kind, which have been prepared in the past by some colleges with traditional ties, assist the scholarship officers to meet and discuss the financial needs of the applicants they have in common before notifying them of awards. The lists prepared by the CSS will include all applicants whose parents submitted financial statements by March 7. Other colleges are expected to prepare their own lists of joint candidates and to hold meetings at which their applications will be considered.

In September, after all awards to the class entering in 1955 have been made, the CSS will send a list to each participating college of all candidates who applied to it for financial aid and ask for a report of the action taken on each candidate's application. These reports will be used to compile lists for each college of the kinds and amounts of aid offered to and accepted by those applicants which it had in common with other participating colleges.

Increased participation: The possibility of opening the Service to colleges which are not members of the College Board and to responsible sponsored scholarship programs will be considered this spring. Ten non-member colleges and a number of organizations have expressed interest in the program. In any case, the number of participating member colleges, now 95 (see page 21), is expected to increase before the beginning of the new academic year in September.

Upperclass form: A special adaptation of the CSS form has been made available to participating colleges at their request for use in annual reviews of upperclass scholarship awards. More than 7,000 copies of the form have been ordered by 30 colleges.

Future possibilities

Computation service: The CSS hopes to be able to assist colleges during the 1955-56 academic year in their computations of the amount of financial aid needed by applicants. Details of this service must await committee study and information which will be yielded by the first year's operation. Possibilities range from such detailed and specific work as actual estimates of the amounts of financial support to be reasonably expected from the families of individual applicants to more general summaries of key information obtained from each financial statement accompanied by tables of normative data relating to a large variety of family situations and financial factors.

Computation manual: A new computation manual to improve on the preliminary edition distributed to participating colleges this year will be prepared during the summer. The manual will be planned to accommodate any changes which may be made in the computation procedure as a result of the colleges' experience with it in 1954-55 and any improvements suggested by data now being collected from the financial statements themselves.